KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 3 урок > MDE, sample size

> Оглавление

Все, о чём говорится в лекции, относится к $i.i.d.$

> Подготовка к эксперименту

Представьте, что вы провели A/B-тест, но результат не оказался статистически значимым.

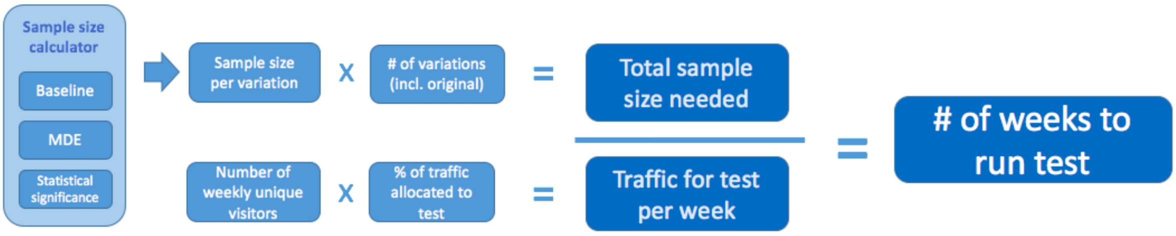
Так могло получиться, если:

1. Эффекта действительно нет;
2. Эффект недостаточно большой, чтобы его можно было обнаружить при текущем дизайне.

> Что нужно посчитать до начала эксперимента

Обязательно заранее необходимо установить следующие параметры:

- Минимальная величина эффекта, которую мы хотим быть способны обнаружить;
- Уровень статистической значимости — вероятности ошибок I и II рода;
- Доля пользователей в эксперименте.



Пример: рассмотрим X5, где пилоты проводятся на магазинах; величины, в которых измеряют выборку — "магазино-дни".

Если получили, что, по некоторым формулам, необходимо 7000 магазино-дней, то есть несколько вариантов:

1. Взять 10 магазинов и проводить пилот 700 дней (долго).
2. Взять 100 магазинов и проводить пилот неделю.

Почему такой подход неверен: каждодневные подсчёты некоторых величин для магазинов — не $i.i.d.$, т.к. продажи магазина X сегодня сильно зависят от продаж вчера. Однако это может быть неплохой аппроксимацией в некоторых случаях.

> Выбор размера: статистика vs риски

Почему мы хотим иметь большой размер эксперимента:

- Чем больше группы, тем они репрезентативнее, т.е. лучше отражают генеральную совокупность.
- Мы получаем меньший разброс, выше статистическая значимость.
- При том же уровне значимости можно быстрее получить результат.

С точки зрения статистики лучше всего поделить всех пользователей 50/50 между экспериментальной и контрольной группами.

Почему мы хотим иметь маленький размер эксперимента:

- Одновременно может идти несколько экспериментов. Мы не хотим, чтобы эксперименты влияли друг на друга.
- Проведение эксперимента может быть затратным.
- Любой эксперимент несёт риски экономических потерь.

Необходимо искать правильный баланс.

> Тестирование гипотез

Изначально выдвигается нулевая гипотеза: группы не отличаются, т.е. предполагаем, что наши усилия не имели эффекта.

Для формализма скажем, что есть две выборки $X = x_1, x_2, ...x_n$ и $Y = y_1, y_2, ...y_n$

$Dx_i = \sigma_X^2$

$Dy_i = \sigma_Y^2$

Была выдвинута гипотеза H_0: $\mu_X = \mu_Y$, H_1: $\mu_Y > \mu_X$

Нужно посчитать вероятность того, что наблюдаемые различия появятся при выполнении нулевой гипотезы. Эта вероятность называется ошибкой I рода.

$$FPR = P(\frac{\overline{Y} - \overline{X}}{\sqrt{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}} > C|H_0) \to \min$$

В результате изменений среднее значение сместилось. Величину смещения мы не знаем, но можем оценить по выборке.

Вероятность верно принять альтернативную гипотезу называется мощностью статистического критерия.

$$Power = P(\frac{\overline{Y} - \overline{X}}{\sqrt{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}} > C|H_1) \to \max$$

> Односторонний и двусторонний тесты

Мы можем выбирать разные постановки тестирования гипотезы:

- Эффект больше пороговой величины;

- Эффект отличается от нуля больше, чем на пороговую величину.

После фиксации α, H_0, H_1 и выбора статистики T можно построить для неё критическую область с заданным уровнем значимости. По выбранному уровню значимости мы можем также определить мощность критерия.

Проблема: невозможно одновременно минимизировать ошибки I и II рода.

Принят следующий подход:

1. Фиксация ошибки I рода, α
2. Минимизация ошибки II рода, β

> Ошибки при принятии решений

Экспериментатор хочет:

- Уменьшить ошибку первого рода;
- Увеличить мощность критерия.

Напомним понятия:

Ошибка I рода: большая величина ошибки I рода означает, что мы часто будем находить эффект при его отсутствии. В результате мы потратим много денег на внедрения, которые ничего не дадут.

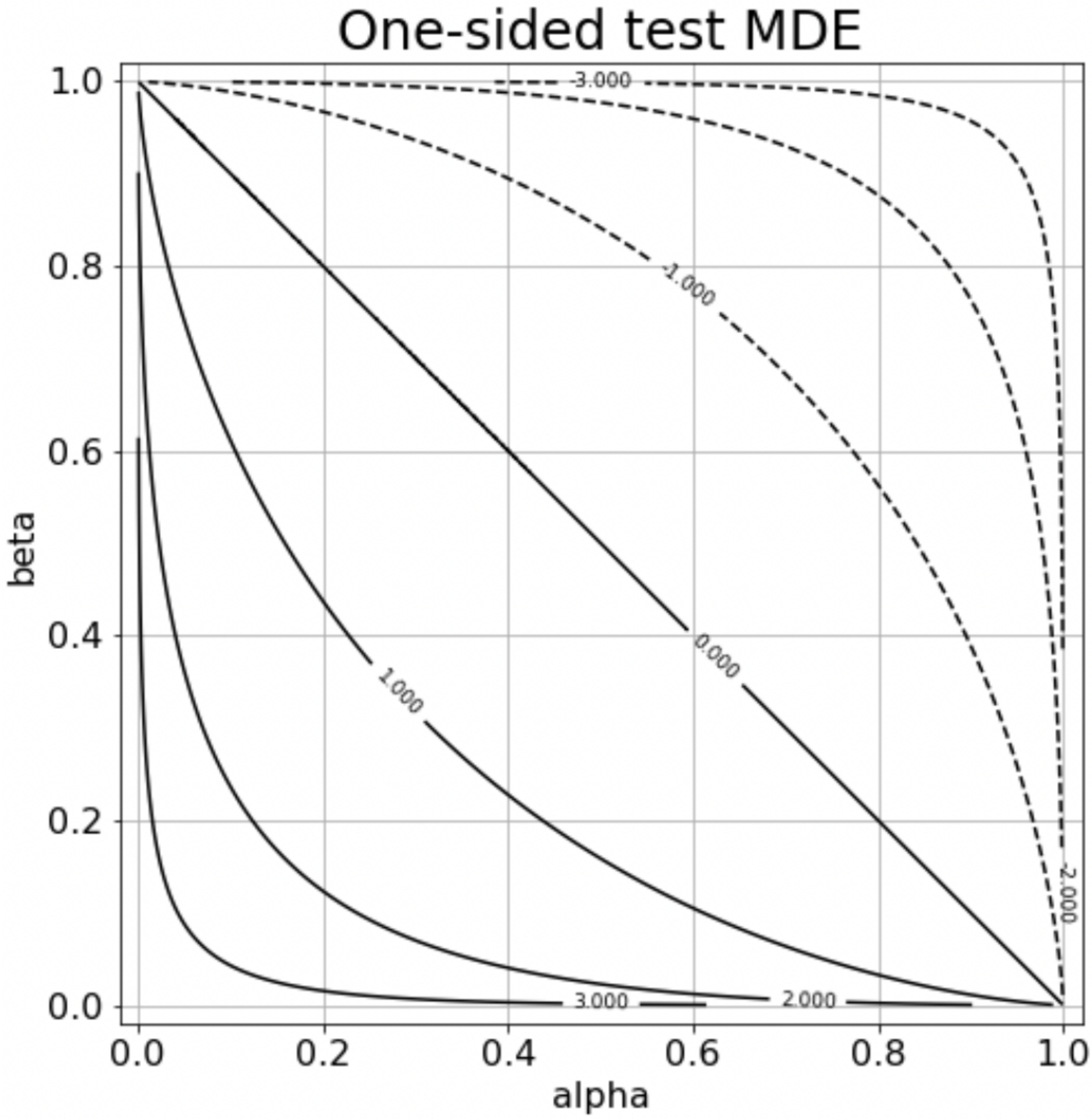
Типичное значение ошибки I рода составляет 5% (уровень значимости 95%).

Мощность критерия: низкая мощность критерия означает, что мы будем часто пропускать позитивные изменения. У нас в руках идея, которая может заработать миллионы, а мы её отвергаем!

Можно выбрать мощность критерия в 80%. Тогда мы обнаружим четыре классные идеи из пяти.

> MDE

MDE — минимальный детектируемый эффект на заданном уровне значимости и с заданной мощностью.



Линии уровня минимально детектируемого эффекта: невозможно одновременно минимизировать ошибки I и II рода — разнонаправленность ошибок

> Математическое обоснование MDE

Статистическая гипотеза: $X_1, X_2, ...X_n \sim N(a, \sigma_0^2)$

Нулевая гипотеза и альтернативная гипотезы: $H_0 : a = a_0, H_1 : a = a_1, a_0 < a_1$

Будем использовать критерий отношения правдоподобия:

$$T(X) = \frac{l_1}{l_0} = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{(X_i - a_1)^2}{2\sigma_0}\}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{(X_i - a_0)^2}{2\sigma_0}\}} = \exp\{\frac{1}{2\sigma_0^2}(\sum_{i=1}^{n} 2X_i(a_1 - a_0) + \sum_{i=1}^{n}(a_0^2 - a_1^2))\}$$

Критерий имеет вид $T(X) \geq c * (\alpha)$ — необходимо найти такой $c * (\alpha)$, чтобы $T(X)$ попадала в него с вероятностью не большей, чем α при условии, что H_0 верна.

$T(X)$ сонаправлена с $\sum_{i=1}^{n} 2X_i$, поэтому перейдём к анализу более простой статистики $\sum_{i=1}^{n} X_i \geq c(\alpha)$.

Хотим зафиксировать вероятность ошибки I рода $P_{H_0}(\sum_{i=1}^{n} X_i \geq c) \leq \alpha$.

Как найти параметр $c(\alpha)$? Воспользуемся ЦПТ.

$$\sum X_i \sim N(na, n\sigma_0^2)$$

$$\frac{\sum X_i - na_0}{\sqrt{n}\sigma_0} \sim N(0,1)$$

$$P(\frac{\sum X_i - na_0}{\sqrt{n}\sigma_0} \geq \frac{c - na_0}{\sqrt{n}\sigma_0}) = \alpha$$

$$1 - \Phi(\frac{c - na_0}{\sqrt{n}\sigma_0}) = \alpha$$

Получаем выражение для границы критической области:

$$c = \Phi - 1(1 - \alpha)\sqrt{n}\sigma_0 + na_0$$

Заметим, что c не зависит от a_1 и верно $\forall a_1 : a_1 > a_0$.

Теперь разберёмся с ошибкой второго рода:

$P_{H_1}(\sum_{i=1}^{n} X_i \geq c) \geq 1 - \beta$ — эффект есть при H_1

$$P_{H_1}(\frac{\sum_{i=1}^{n} X_i - na_1}{\sqrt{n}\sigma_0} \geq \frac{c - na_1}{\sqrt{n}\sigma_0}) \geq 1 - \beta$$

Подставим выражение для c:

$$P_{H_1}(\frac{\sum_{i=1}^{n} X_i - na_1}{\sqrt{n}\sigma_0} \geq \frac{\Phi^{-1}(1 - \alpha)\sqrt{n}\sigma_0 + na_0 - na_1}{\sqrt{n}\sigma_0}) \geq 1 - \beta$$

Воспользуемся ЦПТ:

$$1 - \Phi(\Phi^{-1}(1-\alpha) + \frac{\sqrt{n}(a_0 - a_1)}{\sigma_0}) \geq 1 - \beta$$

$\epsilon = a_1 - a_0 \geq \dfrac{(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta))\sigma_0}{\sqrt{n}}$ — ожидаемый эффект: разница, которую мы хотим быть способны обнаружить.

Покажем, что $\Phi^{-1}(\beta) = -\Phi^{-1}(1-\beta)$

Пусть $\Phi(x) = \beta$

Известно, что $\Phi(x) + \Phi(-x) = 1$, тогда

$\Phi(-x) = 1 - \Phi(x) = 1 - \beta$

$x = \Phi^{-1}(1-\beta)$

С другой стороны: $x = \Phi^{-1}(\beta)$

Подставив, получим доказываемое равенство.

Выпишем результат:

ϵ — размер эффекта.

α — допустимая ошибка I рода.

β — допустимая ошибка II рода.

σ_X^2, σ_Y^2 — дисперсии выборок.

n — размеры выборок.

$$\epsilon^2 > \frac{[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2(\sigma_X^2 + \sigma_Y^2)}{n} | \cdot \frac{n}{\epsilon^2}$$

$$n > \frac{[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2(\sigma_X^2 + \sigma_Y^2)}{\epsilon^2}$$

> Sample size

Оценим минимальный размер выборки, который необходим, чтобы обнаружить ожидаемый эффект при фиксированных ошибках I и II рода:

$$n > \frac{[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2(\sigma_X^2 + \sigma_Y^2)}{\epsilon^2}$$

Рассмотрим зависимость n от параметров:

$\alpha/\beta \downarrow$, то $n \uparrow$

$\sigma_X^2 \downarrow$, то $n \downarrow$

$\epsilon \downarrow$, то $n \uparrow$

$n \sim \frac{1}{\epsilon^2}$

> Variance reduction

В формуле оценки n можем влиять только на $\sigma_X^2 + \sigma_Y^2$

Для снижения дисперсии нужно много данных.

Что в таком случае можно делать:

- Повышать качество собираемых данных;
- Фильтровать выбросы;
- Использовать CUPED.